BAYESIAN CLASSIFICATION

Bayesian Classifiers

Bayesian classifiers are statistical classifiers

They can calculate the probability that a given sample belongs to a particular class

Bayesian classification is based on Bayes theorem

Bayesian classifiers have exhibited high accuracy and speed when applied to large databases

BAYESIAN CLASSIFICATION

Bayes Theorem

Let *X* be a data sample, e.g. red and round fruit

Let *H* be some hypothesis, such as that *X* belongs to a specified class *C* (e.g. X is an apple)

For classification problems, we want to determine $P(H|X)$, the probability that the hypothesis *H* holds given the observed data sample *X*

BAYESIAN CLASSIFICATION

Prior Probability

The probability *P(H)* is called the prior probability of *H*, i.e the probability that any given data sample is an apple, regardless of how the data sample looks

The probability *P(H|X)* is called posterior probability. It is based on more information, then the prior probability *P(H)* which is independent of *X*

BAYESIAN CLASSIFICATION

Posterior Probability

Example:

Let the data samples be fruits described by their colour and shape

Suppose X = red and round

and H = hypothesis that X is an apple

Then $P(H|X)$ is the probability that X is an apple given that X is red and round

BAYESIAN CLASSIFICATION

Bayes Theorem

It provides a way of calculating the posterior probability

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)}$$

***P(X|H)* is the posterior probability of *X* given *H* (it is the probability that *X* is red and round given that *X* is an apple)**

***P(X)* is the prior probability of *X* (probability that a data sample is red and round)**

BAYESIAN CLASSIFICATION

Bayes Theorem: Proof

The posterior probability of the fruit being an apple given that its shape is round and its colour is red is

$$P(H|X) = |H \wedge X| / |X|$$

i.e. the number of apples which are red and round divided by the total number of red and round fruits

Since $P(H \wedge X) = |H \wedge X|$ / |total fruits of all size and shapes|
and $P(X) = |X|$ / |total fruits of all size and shapes|

Hence $P(H|X) = P(H \wedge X) / P(X)$

BAYESIAN CLASSIFICATION

Bayes Theorem: Proof

Similarly $P(X|H) = P(H \wedge X) / P(H)$

Hence we have $P(H \wedge X) = P(H|X)P(X)$
And also $P(H \wedge X) = P(X|H)P(H)$

Therefore $P(H|X)P(X) = P(X|H)P(H)$

And hence $P(H|X) = P(X|H)\, P(H) / P(X)$

BAYESIAN CLASSIFICATION

Naïve (Simple) Bayesian Classification

Studies comparing classification algorithms have found that the simple Bayesian classifier is comparable in performance with decision tree and neural network classifiers

It works as follows:

1. **Each data sample is represented by an n-dimensional feature vector, $X = (x_1, x_2, \ldots, x_n)$, depicting n measurements made on the sample from n attributes, respectively $A_1, A_2, \ldots A_n$**

BAYESIAN CLASSIFICATION

Naïve (Simple) Bayesian Classification

2. Suppose that there are m classes C_1, C_2, … C_m. Given an unknown data sample, X (i.e. having no class label), the classifier will predict that X belongs to the class having the highest posterior probability given X

Thus if $P(C_i|X) > P(C_j|X)$ for $1 \leq j \leq m$, $j \neq i$
then X is assigned to C_i

BAYESIAN CLASSIFICATION

Naïve (Simple) Bayesian Classification

3. As P(X) is constant for all classes, only $P(X|C_i)\ P(C_i)$ needs to be calculated

The class prior probabilities may be estimated by

$$P(C_i) = s_i / s$$

where s_i is the number of training samples of class C_i

& s is the total number of training samples

If class prior probabilities are equal (or not known and thus assumed to be equal) then we need to calculate only $P(X|C_i)$

BAYESIAN CLASSIFICATION

Naïve (Simple) Bayesian Classification

4. Given data sets with many attributes, it would be extremely computationally expensive to compute $P(X|C_i)$

For example, assuming the attributes of colour and shape to be Boolean, we need to store 4 probabilities for the category apple

$$P(\neg red \wedge \neg round \mid apple)$$
$$P(\neg red \wedge round \mid apple)$$
$$P(red \wedge \neg round \mid apple)$$
$$P(red \wedge round \mid apple)$$

If there are 6 attributes and they are Boolean, then we need to store 2^6 probabilities

BAYESIAN CLASSIFICATION

Naïve (Simple) Bayesian Classification

In order to reduce computation, the naïve assumption of *class conditional independence* is made

This presumes that the values of the attributes are conditionally independent of one another, given the class label of the sample (we assume that there are no dependence relationships among the attributes)

BAYESIAN CLASSIFICATION

Naïve (Simple) Bayesian Classification

Thus $P(X|C_i) = \prod^{n}_{k=1} P(x_k|C_i)$

Example

P(colour $\wedge$ shape | apple) = P(colour | apple) P(shape | apple)

For 6 Boolean attributes, we would have only 12 probabilities to store instead of $2^6 = 64$

Similarly for 6, three valued attributes, we would have 18 probabilities to store instead of 3^6

BAYESIAN CLASSIFICATION

Naïve (Simple) Bayesian Classification

The probabilities $P(x_1|C_i)$, $P(x_2|C_i)$, …, $P(x_n|C_i)$ can be estimated from the training samples, where

For an attribute A_k, which can take on the values x_{1k}, x_{2k}, …
e.g. colour = red, green, …

$$P(x_k|C_i) = s_{ik}/s_i$$

where s_{ik} is the number of training samples of class C_i having the value x_k for A_k
and s_i is the number of training samples belonging to C_i

e.g. P(red|apple) = 7/10 if 7 out of 10 apples are red

BAYESIAN CLASSIFICATION

Naïve (Simple) Bayesian Classification

If A_k is continuous valued then the attribute is typically assumed to have a Gaussian distribution so that

$P(x_k|C_i) \quad = g(x_k, \mu_{Ci}, \sigma_{Ci})$

= gaussian (normal) density function for attribute A_k

while μ_{Ci} is the mean
and σ_{Ci} is the standard deviation,
given the values for attribute A_k for class C_i training samples

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) = \frac{1}{\sqrt{2\pi}\sigma_{C_i}} e^{-\frac{(x-\mu_{C_i})^2}{2\sigma_{C_i}^2}},$$

BAYESIAN CLASSIFICATION

Naïve (Simple) Bayesian Classification

rid	age	income	student	credit_rating	Class: buys_computer
1	<30	high	no	fair	no
2	<30	high	no	excellent	no
3	30-40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	30-40	low	yes	excellent	yes
8	<30	medium	no	fair	no
9	<30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<30	medium	yes	excellent	yes
12	30-40	medium	no	excellent	yes
13	30-40	high	yes	fair	yes
14	>40	medium	no	excellent	no

BAYESIAN CLASSIFICATION

Naïve (Simple) Bayesian Classification

Example:

Let C_1 = class buy computer and C_2 = class not buy computer

The unknown sample:
X = {age ≤ 30, income = medium, student = yes, credit-rating = fair}

The prior probability of each class can be computed as

P(buy computer = yes) = 9/14 = 0.643
P(buy_computer = no) = 5/14 = 0.357

BAYESIAN CLASSIFICATION

Naïve (Simple) Bayesian Classification

Example:

To compute P(X|Ci) we compute the following conditional probabilities

$$
\begin{aligned}
&P(age = \text{"<30"} \mid buys\_computer = yes) && = 2/9 = 0.222 \\
&P(age = \text{"<30"} \mid buys\_computer = no) && = 3/5 = 0.600 \\
&P(income = medium \mid buys\_computer = yes) && = 4/9 = 0.444 \\
&P(income = medium \mid buys\_computer = no) && = 2/5 = 0.400 \\
&P(student = yes \mid buys\_computer = yes) && = 6/9 = 0.667 \\
&P(student = yes \mid buys\_computer = no) && = 1/5 = 0.200 \\
&P(credit\_rating = fair \mid buys\_computer = yes) && = 6/9 = 0.667 \\
&P(credit\_rating = fair \mid buys\_computer = no) && = 2/5 = 0.400
\end{aligned}
$$

BAYESIAN CLASSIFICATION

Naïve (Simple) Bayesian Classification

Example:

Using the above probabilities we obtain

$$P(X|buys\_computer = yes) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$$
$$P(X|buys\_computer = no) = 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.019$$

And hence the naïve Bayesian classifier predicts that the student will buy computer, because

$$P(X|buys\_computer = yes)P(buys\_computer = yes) = 0.044 \times 0.643 = 0.028$$
$$P(X|buys\_computer = no)P(buys\_computer = no) = 0.019 \times 0.357 = 0.007$$